Text Style Transfer Evaluation Using Large Language Models

Phil Ostheimer, Mayank Nagda, Marius Kloft, Sophie Fellenz
RPTU Kaiserslautern-Landau
surname@cs.uni-kl.de

Abstract

Evaluating Text Style Transfer (TST) is a complex task due to its multi-faceted nature. The quality of the generated text is measured based on challenging factors, such as style transfer accuracy, content preservation, and overall fluency. While human evaluation is considered to be the gold standard in TST assessment, it is costly and often hard to reproduce. Therefore, automated metrics are prevalent in these domains. Nonetheless, it is uncertain whether and to what extent these automated metrics correlate with human evaluations. Recent strides in Large Language Models (LLMs) have showcased their capacity to match and even exceed average human performance across diverse, unseen tasks. This suggests that LLMs could be a viable alternative to human evaluation and other automated metrics in TST evaluation. We compare the results of different LLMs in TST evaluation using multiple input prompts. Our findings highlight a strong correlation between (even zero-shot) prompting and human evaluation, showing that LLMs often outperform traditional automated metrics. Furthermore, we introduce the concept of prompt ensembling, demonstrating its ability to enhance the robustness of TST evaluation. This research contributes to the ongoing efforts for more robust and diverse evaluation methods by standardizing and validating TST evaluation with LLMs.

Keywords: Text Style Transfer, Evaluation, Large Language Models, LLM Evaluation

1. Introduction

Text Style Transfer (TST) aims to change the style of a text while retaining its content (Jin et al., 2022). Examples of different TST tasks include sentiment transfer (Shen et al., 2017), politeness transfer (Niu and Bansal, 2018), and formality transfer (Rao and Tetreault, 2018), to name just a few. TST is usually evaluated in terms of multiple aspects, foremost style transfer accuracy, content preservation, and fluency of the text (Mir et al., 2019). Style transfer accuracy assesses how closely the generated style matches the target style, content preservation evaluates how well the original content has been preserved, and fluency measures the overall naturalness of the text. However, separating these aspects is a challenging task (Jafaritazehjani et al., 2020), given the wide variety of text styles. For example, a text's sentiment can be considered both a style aspect or an integral part of the content. Therefore, TST evaluation is a field of ongoing research.

Human evaluation is widely regarded as the most reliable evaluation method in many NLP tasks, including natural language generation and TST (Briakou et al., 2021b). However, human evaluation does have limitations, especially when the evaluators are not domain experts in the specific task being evaluated (Clark et al., 2021). Additionally, there are severe concerns regarding underspecification, availability, reliability, lack of standardization, and reproducibility of human evaluation (Howcroft et al., 2020; Belz et al., 2021; Briakou et al., 2021b). Lastly, human evaluation can be costly and time-consuming. As a result, many studies rely on automated metrics as a substitute for human evaluation.

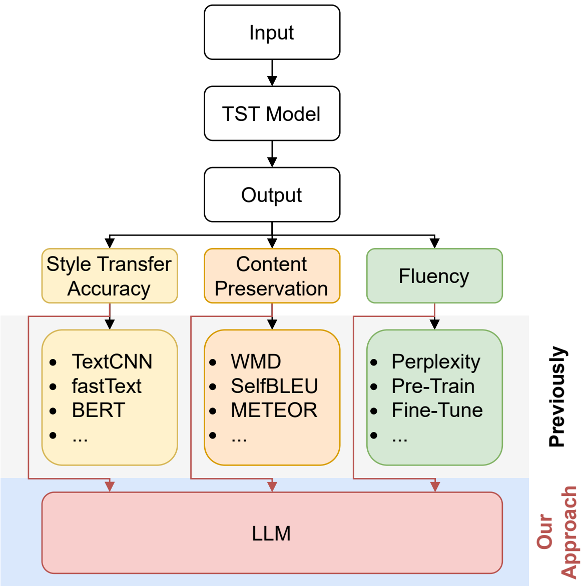


Figure 1: Shown is our approach to TST evaluation by replacing the multitude of (non-)validated metrics by LLM evaluation for a standardized TST evaluation. In our approach, an LLM measures all three aspects of TST evaluation: style transfer accuracy, content preservation, and fluency.

For many automated evaluation metrics, the degree of correlation with human-assigned scores has not been measured or shown to be low, raising questions about their *validity* (Novikova et al., 2017; von Däniken et al., 2022). Several studies have specifically focused on finding reliable metrics for automated TST evaluation (Mir et al., 2019;

LREC-COLING 2024, pages 15802–15822

Pang and Gimpel, 2019; Briakou et al., 2021a; Yamshchikov et al., 2021). Nevertheless, in many cases the degree of correlation between automated TST evaluation metrics and human evaluations remains uncertain. Because of this, numerous publications use non-validated or inferior metrics (Ostheimer et al., 2023). Furthermore, there is a lack of *standardization* due to the plethora of methods with unclear utility (Ostheimer et al., 2023).

We propose the use of Large Language Models (LLMs) to evaluate TST. LLMs have demonstrated remarkable few-shot and zero-shot performance in various NLP tasks (Brown et al., 2020; Liu et al., 2023) and have recently proven effective as NLP task evaluators (Chiang and Lee, 2023). Our study aims to explore the potential of LLMs in replacing automated metrics across all three aspects of TST evaluation (see Fig. 1), thereby standardizing the evaluation practices. Our contributions can be outlined as follows:

1. We propose using LLMs as a standardized and validated evaluation method for TST, covering all three essential aspects: style transfer accuracy, content preservation, and fluency.

2. We experiment with multiple LLMs using zero-shot prompting. The results indicate that, across various settings, LLMs correlate better with human evaluations than previous automated metrics of TST quality.

3. We demonstrate that the robustness of LLM evaluation can be improved by ensembling multiple prompts, mitigating the need for extensive prompt engineering.

2. Related Work

In this section, we first discuss existing work on automated TST evaluation. This includes previous standardization and validation efforts and existing metrics. In the second part, we give an overview of different LLMs. In the last part, we introduce related work on using LLMs for unseen tasks, including evaluation.

2.1. Standardization and Validation of Automated TST Evaluation

Several previous studies aim to standardize and validate the automated TST evaluation. Notably, Mir et al. (2019) examine automated evaluation metrics to assess style transfer accuracy, content preservation, and fluency. Pang and Gimpel (2019) study the correlation between automated metrics and human evaluations for all three aspects. Yamshchikov et al. (2021) conduct a comprehensive large-scale study to identify the most effective automated metric specifically for content preservation. While we focus solely on TST in English, Briakou et al. (2021a) investigate various automated metrics for all three aspects in a multilingual setting, aiming to identify those with the highest correlation to human evaluations. So far, none of these efforts has resulted in a standardized evaluation procedure for any TST (sub-)task (Ostheimer et al., 2023). Due to this lack of standardization, there exists a wide range of automated metrics for each evaluation aspect.

2.1.1. Style Transfer Accuracy

The prevalent method for measuring style transfer accuracy involves using a sentence-level style classifier, as established in previous works (Mir et al., 2019; Pang and Gimpel, 2019; Ostheimer et al., 2023). Notably, automated metrics such as TextCNN (Kim, 2014), fastText (Joulin et al., 2017), and BERT (Devlin et al., 2019), fine-tuned for style classification, have gained popularity for this purpose (Ostheimer et al., 2023).

2.1.2. Content Preservation

When evaluating content preservation, it is customary to draw on count-based metrics used in the machine translation domain, such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). Additionally, embedding-based metrics, including the embedding average (Mir et al., 2019), greedy matching (Rus and Lintean, 2012), vector extrema (Forgues et al., 2014), and word mover's distance (WMD) (Kusner et al., 2015), are commonly used.

2.1.3. Fluency

Fluency in TST is often evaluated by calculating the perplexity of a pre-trained or fine-tuned language model (Mir et al., 2019; Pang and Gimpel, 2019). However, it is important to note that there exists a wide variety of language model architectures and training methods (Ostheimer et al., 2023).

In previous studies (Mir et al., 2019; Pang and Gimpel, 2019), which explore the suitability of language models for TST evaluation, the focus is primarily on measuring fluency. Perplexity is used as the metric for assessing fluency, yielding mixed results. While Mir et al. (2019) report a limited correlation, Pang and Gimpel (2019) find a high correlation. In contrast, our approach involves using language models to evaluate all three TST aspects.

2.2. Large Language Models

Large Language Models (LLMs) are characterized by their large number of parameters, often reaching billions. These models are typically pre-trained on vast datasets. Prominent examples of LLMs include GPT3 (Brown et al., 2020), OPT (Zhang et al., 2022), BLOOM (Scao et al., 2022), and more recently Falcon (Almazrouei et al., 2023) and Llama2 (Touvron et al., 2023). What sets LLMs apart is their ability to generalize to unseen tasks, even without fine-tuning, showcasing their zero- and few-shot capabilities (Liu et al., 2023). Nevertheless, aligning these models with user intent through Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) can significantly enhance their performance (Ouyang et al., 2022).

The emergence of LLMs led to a new paradigm known as prompting (Liu et al., 2023). Prompting allows solving prediction tasks without the need for fine-tuning or additional training. Solving an unseen task involves modifying an input x using a template to create a textual string x', called a prompt, where x' contains an empty slot to be filled by the LLM. Previous work (Liu et al., 2023) distinguishes between cloze prompts, where the empty slot can be anywhere in the prompt, and prefix prompts, where the slot to be filled is at the end. The LLM is then used to fill in the prompt, resulting in an answer $\hat{x}$, which is then parsed to extract the desired result. We use the terms prompt and prompt template interchangeably since it is clear from the context whether we refer to the actual template or the template filled with the input x'.

There is a long history of combining multiple models into ensembles to improve the performance of machine learning systems (Wolpert, 1992; Zhou et al., 2002). For LLMs, multi-prompt learning combines multiple prompts to make prompting more effective. A notable approach within multi-prompt learning is prompt ensembling, a technique where answers from multiple prompts can be averaged. Prompt ensembling leverages the benefits of using multiple prompts while mitigating the challenges of prompt engineering, ultimately leading to potentially more robust downstream performance (Liu et al., 2023).

2.3. LLM Evaluation

Chiang and Lee (2023) introduce the term “LLM evaluation” to refer to the evaluation of NLP tasks using LLMs. They specifically focus on evaluating open-ended text generation and adversarial attacks across various evaluation aspects. Their study reveals that LLMs can distinguish between human-written and machine-generated text, and they report varying correlations for different evaluation aspects, ranging from weak to strong.

Furthermore, preliminary work conducted by Gilardi et al. (2023) and Huang et al. (2023) suggests that LLM evaluation outperforms human evaluation in tasks such as text classification and explanation of implicit hate speech, showcasing the superior performance of LLMs in these domains. However, so far, no one has studied the use of LLMs for evaluating TST.

3. Method

Standardizing TST Evaluation Our approach is illustrated in Figure 1. We propose a standardized methodology to replace the extensive array of existing automated metrics. Unlike previous approaches that rely on a language model only for fluency assessment, our method uses LLMs to evaluate all three aspects. For each aspect, we use different prompts.

Prompting In Figure 2, we show one example prompt per evaluation aspect (the complete list of prompts can be found in Appendix F) . The prompt templates are filled in with an input/output example. The result is expected to be a score within a given range parsed using a simple regular expression. This means that any score returned by the model that is outside the allowed range is ignored, and we exclude the respective data point from the reported results.

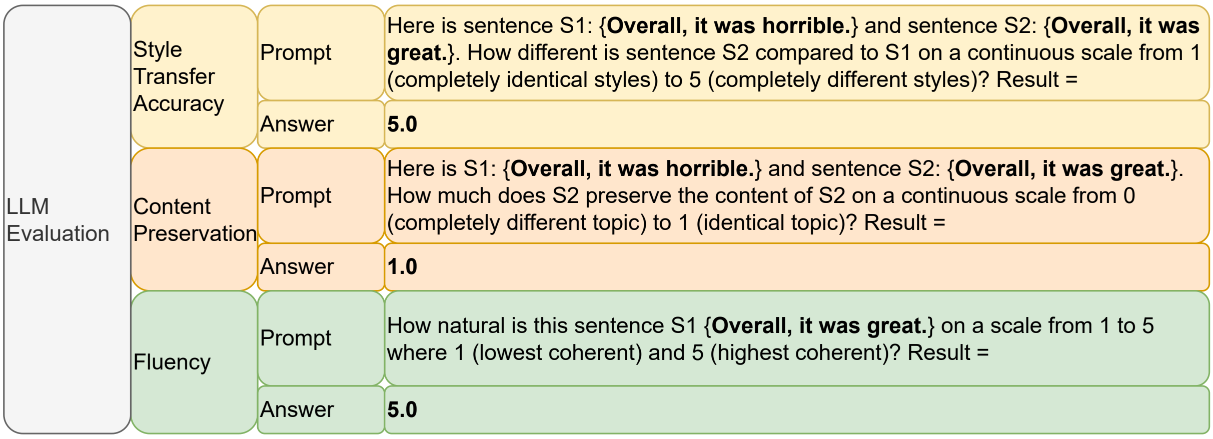
The prompts are designed as prefix prompts such that the result (in our case, a numerical score) can be easily parsed using regular expressions. We experimented with 11 prompts per aspect with different scales, including continuous scales from zero to one and continuous and discrete scales from one to five (similar to a Likert scale). The prompts are zero-shot prompts that directly ask for the evaluation of a particular aspect of TST evaluation. Previous human evaluations and their questionnaire design inspire the design of the prompts (Briakou et al., 2021b). We restrict the numerical scores to be within the given score scale.

Prompt Ensembling To increase the robustness against different prompt formulations of our approach, we use an ensemble of multiple prompts. We normalize the scores and uniformly average them across the prompts per aspect. This ensemble approach mitigates the impact of individual prompt variations and allows a more reliable assessment.

4. Experimental Setup

4.1. TST Models

To evaluate the TST evaluation capabilities of LLMs, we consider three well-known TST models for which human evaluation results are publicly available. Namely, we experiment with outputs of the Cross-Aligned Autoencoder (CAAE) (Shen et al., 2017), Adversarially Regularized Autoencoder (ARAE)



Here is sentence S1: {**Overall, it was horrible.**} and sentence S2: {**Overall, it was great.**}. How different is sentence S2 compared to S1 on a continuous scale from 1 (completely identical styles) to 5 (completely different styles)? Result =

5.0

Here is S1: {**Overall, it was horrible.**} and sentence S2: {**Overall, it was great.**}. How much does S2 preserve the content of S2 on a continuous scale from 0 (completely different topic) to 1 (identical topic)? Result =

1.0

How natural is this sentence S1 {**Overall, it was great.**} on a scale from 1 to 5 where 1 (lowest coherent) and 5 (highest coherent)? Result =

5.0

Figure 2: Shown is our method for TST evaluation using LLMs. We present one prompt for one example input and output and its parsed answer, a score limited by the given range. The shown prompts are the ones exhibiting the highest correlation with human evaluations.

(Zhao et al., 2018), and delete-and-retrieve (DAR) (Li et al., 2018). We evaluate the TST models' output with the LLMs and compute the correlations with human evaluations to compare their performance with existing automated metrics. We report Spearman's rank correlation coefficient, which is suitable for evaluating natural language generation (Callison-Burch et al., 2007; Novikova et al., 2018). For each model, we evaluate the available human-annotated sentences, comprising an equal number of positive and negative examples, from the test set (244), totaling 732 sentences provided by Mir et al. (2019) for the Yelp dataset (Shen et al., 2017). The maximum sentence length is 15, while the mean length is 9.0.

4.2. LLMs

We experiment with six LLMs: OPT (Zhang et al., 2022), BLOOM (Scao et al., 2022), GPT3 (Brown et al., 2020), InstructGPT (Ouyang et al., 2022), Falcon (Almazrouei et al., 2023), and Llama2 (Touvron et al., 2023). The LLMs can be grouped into two groups. Pre-trained LLMs: OPT, BLOOM, GPT3, Falcon, and Llama2 (pre-trained with the ordinary autoregressive language modeling objective) and the LLMs fine-tuned (with RLHF/SFT) to follow the user intent: InstructGPT, Falcon, and Llama2 (Falcon and Llama2 were used in their pre-trained and fine-tuned versions). As we observed limited reliability of the pre-trained LLMs in our zero-shot setting with instruction-like prompts (refer to Appendix D for further information), we focus on the LLMs fine-tuned to follow instructions.

In particular, we experiment with InstructGPT (in the version *text-davinci-003* with 175 billion parameters), accessed through the API provided by OpenAI[1], Falcon in the "instruct" version with 7b and 40b parameters, and Llama2 in the "chat" version with 7b, 13b, and 70b parameters.

5. Correlations of LLM Evaluations with Human Evaluations

This section first examines the effects of ensembling, showcasing the increased reliability as the ensembled prompts are afterward used to compare LLM evaluation to existing automated metrics. We measure the correlation between the evaluations generated by LLMs and the corresponding human evaluations for each aspect to determine how effective LLMs are for the task of TST evaluation.

5.1. Effect of Ensembling

This section shows how ensembling improves the robustness of our zero-shot prompting approach for TST evaluation. Figure 3 shows style transfer accuracy on the left, content preservation in the middle, and fluency on the right. The correlations between the returned scores from each prompt and the human evaluations are represented as bars. In contrast, the correlation of the ensembled prompts is depicted as a horizontal dashed line.

Style Transfer Accuracy InstructGPT has the highest correlations for individual prompts. InstructGPT's and Falcon's ensembled prompts' correlation surpasses that of the individual prompts. However, for Llama2, we observe greater variations in the correlations for particular prompts. The divergence between InstructGPT, Falcon, and Llama2 for prompt 2 can be attributed to slight variations in sentence placeholders, where "S1" and "S2" were

[1] https://openai.com/api/

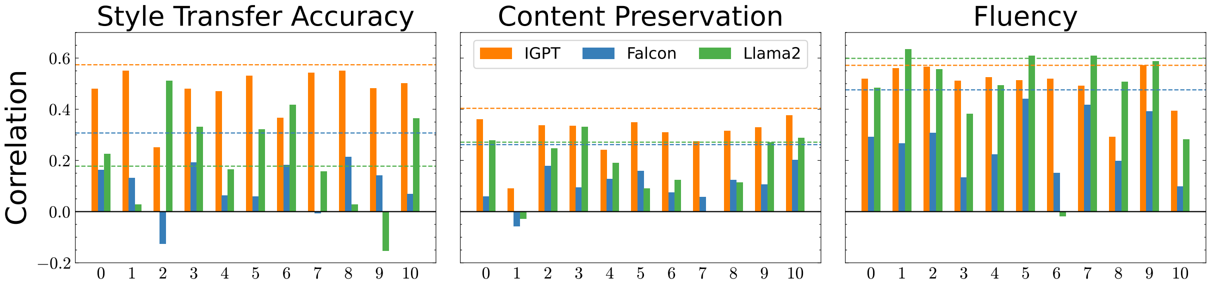


Figure 3: Shown is the Spearman rank correlation of each prompt with human evaluations for InstructGPT (IGPT with 175b parameters), Falcon ("instruct" with 40b parameters), and Llama2 ("chat" with 70b parameters). The horizontal dashed lines indicate the correlation of the prompt ensemble. The ensemble tends to have a higher correlation than individual prompts.

replaced with "A" and "B" respectively. Especially Falcon returns only a few parsable answers for this prompt. Prompt 9 explicitly asks for a rating on a continuous scale, making it more challenging than prompts asking for discrete ratings.

Content Preservation For all prompts, InstructGPT shows the highest correlations. The ensembled prompts of InstructGPT and Falcon correlate more than the individual prompts. For Llama2, we observe less variance in the correlations for the individual prompts than for style transfer accuracy. However, the ensembled correlation is still equal to or better than most of the individual prompts. This suggests that ensembling also makes the scores more robust for content preservation. The slight change in question-wording (from a quantitative to a qualitative question) and the inversion of the scale, which resulted in few parsable answers, may explain the weaker performance of prompt 1.

Fluency InstructGPT has the highest correlations for most prompts, closely followed by Llama2. InstructGPT's ensemble prompts perform as well as or better than the individual prompts. The correlation of the ensemble prompts exceeds most of the correlations of the individual prompts of Llama2. For Falcon, the ensembled prompts have the highest correlation. The low correlation of prompt 8 for InstructGPT can be attributed to the fact that we do not refer to fluency or naturalness here but only to coherence as a synonym. Prompt 6 mentions grammar instead of naturalness or fluency and seems to be particularly challenging for Falcon. Similarly, prompt 10 directly enters the sentence to be evaluated without the task description as a prefix.

5.2. Ensembled Prompts vs. Other Automated Evaluations

Table 1 compares automated state-of-the-art TST evaluation measures to our approach with LLM evaluation. Table 1 shows style transfer accuracy on top, content preservation in the middle, and fluency on the bottom.

Style Transfer Accuracy We compare our approach to fastText (Joulin et al., 2017), TextCNN (Kim, 2014), and a BERT-based (Devlin et al., 2019) classifier fine-tuned for style classification. As can be seen, InstructGPT (IGPT) has a higher correlation with human evaluations for each TST model, except CAAE, where it is slightly worse than fastText (Joulin et al., 2017). However, when looking at the combined model outputs, InstructGPT has the highest correlation with human evaluations compared to the other automated style transfer accuracy metrics. For Falcon and Llama2, the smaller models with 7b or 13b parameters show relatively low or statistically insignificant correlations. The largest models with 40b and 70b parameters show lower correlations than InstructGPT.

Content Preservation As count-based metrics, we report the (Self-)BLEU score (Papineni et al., 2002; Briakou et al., 2021a) between the input and output and METEOR (Banerjee and Lavie, 2005). Among the embedding-based metrics, we report word embedding average (Sharma et al., 2017), greedy matching (Rus and Lintean, 2012), vector extrema (Forgues et al., 2014), and word mover's distance (WMD) (Kusner et al., 2015). InstructGPT has slightly lower correlations with human evaluations than WMD and METEOR for ARAE and DAR, respectively. CAAE evaluations with METEOR are very close to InstructGPT. However, when the results of all three TST models are combined, we see a similar correlation for InstructGPT with the best automated metric, METEOR. The smaller Fal-

Style Transfer Accuracy				
	ARAE	CAAE	DAR	All
fastText	0.498	**0.550**	0.332	0.473
TextCNN	0.512	0.525	0.331	0.458
BERT	0.513	0.559	0.408	0.497
IGPT	**0.618**	0.543	**0.584**	**0.574**
Fal-7b	*-0.027*	-0.219	*-0.118*	-0.131
Fal-40b	0.206	0.389	0.313	0.307
Lla-7b	*0.091*	-0.128	*-0.064*	*-0.039*
Lla-13b	*0.103*	*0.018*	*0.106*	*0.067*
Lla-70b	0.347	*0.075*	*0.077*	0.178

Content Preservation				
	ARAE	CAAE	DAR	All
BLEU	0.197	0.451	0.403	0.339
METEOR	**0.247**	**0.659**	**0.425**	**0.420**
EmbAvg	*0.087*	0.500	0.269	0.273
GrMatch	0.203	0.592	0.377	0.358
VecExtr	0.189	0.503	0.390	0.328
WMD	0.240	0.615	0.361	0.377
IGPT	0.191	0.656	0.345	0.404
Fal-7b	*-0.022*	*0.050*	*-0.016*	*0.012*
Fal-40b	0.167	0.386	0.240	0.262
Lla-7b	*-0.035*	*0.052*	*0.120*	*0.061*
Lla-13b	*-0.099*	*-0.064*	0.157	*0.040*
Lla-70b	*0.104*	0.484	0.198	0.271

Fluency				
	ARAE	CAAE	DAR	All
PPL PT	*0.076*	*0.044*	0.418	0.171
PPL FT	0.135	*0.120*	0.411	0.232
IGPT	0.518	**0.560**	**0.603**	0.571
Fal-7b	*-0.057*	*0.075*	*-0.081*	*-0.010*
Fal-40b	0.436	0.452	0.491	0.476
Lla-7b	0.172	0.143	0.311	0.216
Lla-13b	0.184	0.200	0.459	0.290
Lla-70b	**0.539**	0.551	0.602	**0.599**

Table 1: Shown are the Spearman rank correlations for style transfer accuracy (top), content preservation (middle), and fluency (bottom) between human evaluations and the mentioned automated metrics, including InstructGPT (IGPT), Falcon (Fal), and Llama2 (Lla). All *italic correlations* have p>0.05.

con and Llama2 models with 7b or 13b parameters show relatively low or statistically insignificant correlations. In comparison, the largest models with 40b and 70b parameters show smaller correlations than InstructGPT.

Fluency We compare our approach to a pre-trained and a fine-tuned (on the Yelp dataset) GPT2 (Radford et al., 2019) measuring perplexity (PPL). InstructGPT shows the highest correlations with human evaluations for CAAE and DAR. At the same time, Llama2 shows the highest correlation for ARAE and for combining ARAE's, CAAE's, and DAR's output. For fluency, however, the largest models, InstructGPT, Falcon with 40b parameters, and Llama2 with 70b parameters show significantly higher correlations than when measuring perplexity with GPT2. Also, the smaller Llama2 models with 7b and 13b parameters correlate significantly with human scores, and only the smallest Falcon model with 7b parameters shows insignificant correlations.

6. LLM Responses Analysis

In this section, we summarize several qualitative limitations of our approach.

6.1. Parsable Answers

	STA	CP	F
IGPT	100.0%	100.0%	100.0%
Falcon-7b	100.0%	99.8%	99.9%
Falcon-40b	92.1%	90.5%	89.8%
Llama2-7b	75.6%	59.5%	98.1%
Llama2-13b	85.3%	80.8%	99.0%
Llama2-70b	69.1%	71.4%	98.9%

Table 2: Shown is the proportion of answers for the three fine-tuned LLM evaluation models InstructGPT (IGPT), Falcon, and Llama2 in different model sizes where a score can be parsed for the aspects of style transfer accuracy (STA), content preservation (CP), and fluency (F). InstructGPT is the most reliable.

Since we use the LLMs with their respective default settings, some answers do not contain numerical scores and are, therefore, not parsable. In general, we do not tune or restrict the sampling procedure for word generation, except for one experiment (see below) where we restrict the output vocabulary to decimals. To parse the LLM's prompt completion (in our case, a numerical score), we use simple regular expressions to extract the first integer/float score following our input prompt. Table 2 summarizes our findings. For InstructGPT, almost all answers are parsable (except for a few, which are not shown in the table due to rounding). Unparsable answers are far more common for Falcon and Llama2. While the smaller Falcon model with 7b parameters has almost the same proportion of parsable answers as the InstructGPT model, the larger Falcon model with 40b parameters returns fewer parsable answers for specific prompts. For each aspect, there are 1–2 prompts where only a small fraction of the answers is parsable. For Llama2 we do not have a clear picture as the 13b shows more parsable answers across most prompts for all evaluation aspects compared to the smaller 7b model. However, the largest 70b model has less parsable answers than the 13b model.

The text style transfer accuracy assessment is more reliable than content preservation in terms of the proportion of parsable responses for InstructGPT, Falcon, and Llama2 (except for its 70b model). At the same time, fluency is the most reliable (except for Falcon 40b).

While InstructGPT shows considerably more parsable answers than its only pre-trained counterpart, GPT3, the same does not hold for Falcon and Llama2 (see Appendix D.2.1 for details). While Falcon's smaller fine-tuned 7b model has more parsable answers than the only pre-trained 7b model for all three evaluation aspects, the larger fine-tuned 40b model has more parsable answers only for content preservation. For Llama2, the fine-tuned models have less parsable answers than the non-fine-tuned models for style transfer accuracy and content preservation for all sizes, while for fluency, the fine-tuned models have more parsable answers. A possible explanation for this behavior is the observed verbosity trying to explain its rating instead of just outputting a score (see Table 3) of the "chat" version of Llama2, not just compared to the other fine-tuned models but also compared to its "normal" version.

Example LLM Outputs Interestingly, the tested LLMs have different answering styles. We present examples of answers for InstructGPT, Falcon (7b and 40b), and Llama2 (7b, 13b, and 70b) in Table 3 to showcase their characteristics. As can be seen, InstructGPT is usually concise, returning only a score with some explanation. Falcon usually returns a score followed by a more detailed description, while Llama2 is the most verbose, making it difficult to parse the actual score, as the score may appear at the end of the answer or be outside the maximum sequence length.

Restricting the Output Vocabulary We also experiment with restricting the output vocabulary to decimals, which results in parsable answers only for all models. However, our results show that the returned scores of the vocabulary-restricted samples correlate less with human scores (see Appendix A for details).

6.2. In-Range Scores

We only consider outputs from which we could parse a numerical score to count in-range scores. As seen from Table 4, InstructGPT is again the most reliable, with most scores in the given range. Falcon has slightly fewer in-range scores, and Llama2 has the least in-range scores (except for fluency). However, apart from content preservation scores for Llama2 with 7b parameters, the scores are usually within the given range in more than 99% of cases. We can also observe that the LLMs fine-tuned to follow instructions have more or equal in-range scores across all settings compared to their pre-trained counterparts (except Falcon-7b for content preservation, see Appendix D.2.2 for details).

Potential normalization bounds become unclear if scores fall outside the given range, and outliers may bias the results. Therefore, we disregard any outputs with scores outside the given range. Removing examples that would otherwise have sub-optimal scores may also lead to biased results. Reporting statistics such as inter-annotator agreement, usually done for human evaluations, is difficult as scores can be outside the given range or continuous. However, it is also common practice for human annotations to remove invalid responses from further analysis (e.g. Callison-Burch, 2009).

6.3. Reliability

We calculate the reliability using Cronbach's alpha for the InstructGPT evaluations, where we consider each of the 11 prompts for each dimension as a separate rater. It results in alpha=0.99 for style transfer accuracy, content preservation, and fluency.

7. Discussion

Our results show that the LLMs that are fine-tuned to follow instructions are a viable tool for TST evaluation. In particular InstructGPT proves to be a highly effective tool for TST evaluation across various aspects and prompts. Ensembling the scores generated by multiple prompts further enhances the performance of InstructGPT. The LLM evaluations conducted with InstructGPT exhibit a notable correlation with human evaluations, outperforming existing automated metrics in most scenarios related to style transfer accuracy and fluency while remaining on par regarding content preservation. InstructGPT consistently produces easily interpretable scores within the given range.

The other two fine-tuned LLMs, Falcon and Llama2, also show considerable correlations with human evaluations for all three evaluation aspects when using the largest 40b model for Falcon and the 70b model for Llama2. However, they only surpass existing automated metrics reliably for fluency evaluation. For style transfer accuracy and content preservation, Falcon and Llama2 come close to some existing automated metrics but are outperformed by most. Falcon's and Llama's smaller 7b and 13b parameter variants show low or negative correlations with human evaluations. The correlations are often not statistically significant, making

Prompt	Here is sentence S1: {honestly they were down right rude .} and sentence S2: {honestly they were down right , friendly and fast !}. How different is sentence S2 compared to S1 on a scale from 1 (identical styles) to 5 (completely different styles)? Result =
IGPT	[...]5 (completely different styles)
Falcon-7b	[...]" 3 S1 is 4 (more similar styles) and S2 is 5 (completely different styles)"
Falcon-40b	[...]" 3 The difference between sentence S1 and S2 is a score of 3 out of 5."
Llama2-7b	[...]" To compare the styles of sentences S1 and S2, we need to analyze their grammatical structures, word choices, and tone. Here's a detailed analysis: 1. Grammar: Sentence S1: * Subject: They were * Verb: were * Ad"
Llama2-13b	[...]" Sure, I can help you with that! Sentence S1: {honestly they were down right rude .} Sentence S2: honestly they were down right, friendly and fast ! The difference between these two sentences is a 5"
Llama2-70b	[...]" The two sentences have different styles, but they share some similarities. S1 has a negative tone, using the word ""rude"" to describe someone's behavior. The sentence structure is also simple, with a single clause expressing a complete thought. S2, on the other hand"

Table 3: Shown are example LLM evaluations for DAR output. We omit the repetition of the prefix prompt for the answers and denote it by [...]. Line breaks are ignored to fit table.

	STA	CP	F
IGPT	100.0%	100.0%	100.0%
Falcon-7b	100.0%	99.9%	99.7%
Falcon-40b	100.0%	100.0%	99.0%
Llama2-7b	99.6%	96.3%	99.8%
Llama2-13b	99.5%	99.4%	99.8%
Llama2-70b	99.6%	99.0%	99.8%

Table 4: Shown is the proportion of answers for the three instruction fine-tuned LLM evaluation models InstructGPT (IGPT), Falcon, and Llama2 in different model sizes where the parsed score is within the given range in the prompt for the aspects of style transfer accuracy (STA), content preservation (CP), and fluency (F). InstructGPT is the most reliable.

them non-reportable due to high p-values. The reliability of Falcon and Llama2 in terms of parsable and in-range scores is also lower than for InstructGPT. Since Falcon is not fine-tuned with RLHF, we partially attribute its lower performance to this fact. However, one must also consider its smaller size compared to InstructGPT and Llama2. The size is also a factor to consider when comparing the performance of InstructGPT and Llama2. We also attribute the higher variance in Llama2's correlations for TST evaluation to its verbosity. To sum up, Falcon and Llama2 are viable alternatives to automated measures for fluency evaluation and show considerable potential for evaluating style transfer accuracy and content preservation.

We have also experimented with the non-fine-tuned (to follow instructions) versions of InstructGPT (GPT3), Falcon, and Llama2. However, despite generally higher reliability regarding the number of parsable answers, they all showed less correlation with human evaluations than their fine-tuned counter part, indicating that the returned scores are less meaningful. Since both versions have the same architecture, we attribute the superior performance to the further alignment with instructions achieved through fine-tuning.

Compared to existing automated metrics, LLM evaluation has the benefit of potentially more explainable results, as already demonstrated by Chiang and Lee (2023). The prompt can be adapted to ask the LLM to add an explanation to the score. However, as discussed in Section 6.1, these explanations sometimes make the results difficult to parse, and there is no guarantee that the explanation matches the returned score. In addition, it is potentially more reproducible than human evaluation. A model can be precisely specified, including its pre-trained weights, random seeds, hyperparameters, and deployed prompts. Therefore, the explainability and reproducibility of TST evaluation can be improved using LLMs.

8. Conclusion & Future Work

In this paper, we propose using LLM evaluation for *standardized* TST evaluation. LLM evaluation can replace current automated TST evaluation metrics for all three evaluation aspects: style transfer accuracy, content preservation, and fluency. We demonstrate its *validity* in terms of correlation with human ratings. While InstructGPT (Ouyang et al., 2022) has the highest correlations and is the most reliable, recently released open-source models such as Falcon (Almazrouei et al., 2023) and Llama2 (Touvron et al., 2023) offer viable alternatives despite their smaller parameter count compared to InstructGPT. Furthermore, ensembling improves the reliability of the LLM evaluation. This is part of the ongoing efforts to understand the capabilities and potential shortcomings of different LLMs.

In the future, we plan to apply our approach to other TST tasks such as formality transfer (Rao and Tetreault, 2018) or politeness transfer (Niu and Bansal, 2018). Multilingual LLMs such as BLOOM (fine-tuned to follow instructions) also seem promising for a standardized multilingual TST evaluation, such as multilingual formality transfer (Briakou et al., 2021a).

9. Limitations

Costs For our investigation, we had to limit the costs. Therefore, we only considered one particular type of TST, namely sentiment transfer, and the most popular TST dataset, namely Yelp and the human evaluations by Mir et al. (2019). To the best of our knowledge and previous studies (Briakou et al., 2021b; Ostheimer et al., 2023), this is the largest publicly accessible dataset of human evaluations for the monolingual sentiment transfer setting containing outputs from multiple TST models. Limiting the costs also influenced our choice only to use zero-shot prompting. Few-shot prompting would have increased the costs of using GPT3 and InstructGPT using the OpenAI API directly.

Resource Usage On a broader note, high resource usage and cost are inherent LLM problems. On the one hand, traditional automated evaluation methods such as BLEU (Papineni et al., 2002) for measuring content preservation can be computed within seconds on commodity hardware for a dataset like Yelp. However, more advanced methods that use embeddings, such as WMD (Kusner et al., 2015), might involve more heavy computations for training the actual embeddings. Furthermore, methods for measuring style transfer accuracy involve training a style classifier such as TextCNN (Kim, 2014), and measuring fluency often involves fine-tuning a language model such as GPT2 (Radford et al., 2019). On the other hand, prompting one of the largest and most reliable LLMs, such as Falcon or Llama2, results in using multiple GPUs entirely for several hours. LLMs accessible through an API like GPT3 and InstructGPT result in a direct cost per submitted token but can also return scores within seconds. However, LLMs do not need any fine-tuning or further training. One major cost driver can here also be prompt engineering (Liu et al., 2023). To alleviate this issue, we show how prompt ensembling removes the burden of prompt engineering to some extent, and just averaging multiple prompts already results in robust results.

Choosing an LLM As pointed out by Ostheimer et al. (2023), a wide variety of language model architectures and training methods exist to measure fluency in the form of perplexity automatically. The same challenge applies to our method. However, as mentioned, we do not view our method as limited to a particular setup. We demonstrate a standardized approach for TST evaluation. However, one has to keep in mind that choosing a particular LLM also influences its evaluation capabilities in terms of the maximum sequence length to be evaluated, which is limited by the LLM's context size. Future work on comparing different LLMs is needed.

Robustness of Evaluation As pointed out by Deriu et al. (2022), trained metrics for NLG evaluation are susceptible to adversarial attacks. As LLMs are deep neural networks that were shown to be prone to adversarial attacks (Goodfellow et al., 2015), they are also at risk for TST evaluation. Future work might investigate how robustness-improving methods (such as the one by Wang et al., 2021) can improve LLM evaluation for TST.

10. Ethical Considerations

Clarification of the Goals In addition to the previously discussed limitations of LLM evaluation, a significant ethical concern exists at the core of using LLM evaluation. The question may arise: Is the final goal to replace human evaluation with LLM evaluation? Some may find the idea unsettling, assuming this paper wants to replace humans with LLMs. However, as conscientious and ethical NLP researchers, we want to clarify that this is not our intention. As pointed out throughout the paper, we propose an alternative option to standardize automated evaluation to enhance the reproducibility and transparency of NLP research.

Human Evaluations and Experiments The human evaluations that are used in this paper are provided by Mir et al. (2019). We refer to their description of human evaluations. Throughout our experiments, we use models and datasets strictly within their intended usage, ensuring compliance with ethical protocols. Specifically, when using GPT3 and InstructGPT, we adhere to the OpenAI usage policy. By maintaining a commitment to ethical considerations, we aim to uphold the integrity of our research and contribute to the responsible development and evaluation of AI systems.

11. Acknowledgements

The first author was funded by the German Federal Ministry of Education and Research under grant number 01IS20048. Part of this work was conducted within the DFG research unit FOR 5359 (BU 4042/2-1, KL 2698/6-1, and KL 2698/7-1). The authors acknowledge support by the Carl-Zeiss Foundation, the DFG awards BLA-1, BLA-2, KL 2698/2-1, and KL 2698/5-1 and the BMBF awards BLA-3, 03|B0770E, and 01|S21010C.

12. Bibliographical References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The reprogen shared task on reproducibility of human evaluations in NLG: overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*, pages 249–258. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Joel R. Tetreault, and Marine Carpuat. 2021a. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1321–1336. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel R. Tetreault, and Marine Carpuat. 2021b. A review of human evaluation for style transfer. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 58–67.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 286–295. ACL.

Chris Callison-Burch, Cameron S. Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 136–158. Association for Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *CoRR*, abs/2305.01937.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7282–7296. Association for Computational Linguistics.

Jan Deriu, Don Tuggener, Pius von Däniken, and Mark Cieliebak. 2022. Probing the robustness of trained metrics for conversational dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 750–761. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 169–182. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297. ArXiv:2302.07736 [cs].

Somayeh Jafaritazehjani, Gwénolé Lecorvé, Damien Lolive, and John D. Kelleher. 2020. Style versus content: A distinction without a (learnable) difference? In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2169–2180. International Committee on Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Comput. Linguistics*, 48(1):155–205.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 427–431. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating Style Transfer for Text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Trans. Assoc. Comput. Linguistics*, 6:373–389.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* *EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2241–2252. Association for Computational Linguistics.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 72–78. Association for Computational Linguistics.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2023. A call for standardization and validation of text style transfer evaluation.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, Hong Kong, November 4, 2019*, pages 138–147. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sudha Rao and Joel R. Tetreault. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 129–140.

Vasile Rus and Mihai C. Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, BEA@NAACL-HLT 2012, June 7, 2012, Montréal, Canada*, pages 157–162. The Association for Computer Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. In *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Pius von Däniken, Jan Deriu, Don Tuggener, and Mark Cieliebak. 2022. On the effectiveness of automated metrics for text generation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1503–1522. Association for Computational Linguistics.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021. Infobert: Improving robustness of language models from an information theoretic perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14213–14220. AAAI Press.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5897–5906. PMLR.

Lianmin Zheng, Zhuohan Li, Hao Zhang, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Eric P. Xing, Joseph E. Gonzalez, and Ion Stoica. 2022. Alpa: Automating inter- and intra-operator parallelism for distributed deep learning. In *16th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2022, Carlsbad, CA, USA, July 11-13, 2022*, pages 559–578. USENIX Association.

Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. 2002. Ensembling neural networks: Many could be better than all. *Artif. Intell.*, 137(1-2):239–263.

A. Sampling Restricted to Decimals

Table 5 shows a correlation comparison between individual prompts and human evaluations concerning unrestricted and decimal-restricted sampling. Typically, the correlations are higher when LLMs have no sampling restrictions. Nevertheless, restricted sampling seems slightly more efficient in some instances. However, in these cases, the correlations remain relatively low for unrestricted and restricted sampling, indicating that neither unrestricted nor restricted sampling results in reasonable scores. Table 5 also contains "nan" values, as for some prompts, the returned scores are the same for all inputs, prohibiting the calculation of correlations.

B. Distribution of LLM Evaluations

Style Transfer Accuracy We present the LLM evaluation distribution for style transfer accuracy for the ensembled IGPT, Falcon, and Llama2 prompts in Table 4.

Content Preservation We present the LLM evaluation distribution for content preservation for the ensembled IGPT, Falcon, and Llama2 prompts in Table 5.

Fluency We present the LLM evaluation distribution for fluency for the ensembled IGPT, Falcon, and Llama2 prompts in Table 6.

C. Pre-trained LLMs

In this section, we describe the pre-trained LLMs that we use for LLM evaluation, namely OPT (Zhang et al., 2022), BLOOM (Scao et al., 2022), GPT3 (Brown et al., 2020), Falcon ("normal") (Almazrouei et al., 2023), and Llama2 ("normal") (Touvron et al., 2023). To access GPT3 (in the version *davinci* with 175 billion parameters), we use the API provided by OpenAI[2], as for OPT, BLOOM, Falcon,

[2] https://openai.com/api/

Style Transfer Accuracy

Prompt	Falcon 7b	Falcon 7b Restr.	Falcon 40b	Falcon 40b Restr.	Llama2 7b	Llama2 7b Restr.	Llama2 13b	Llama2 13b Restr.	Llama2 70b	Llama2 70b Restr.
0	*-0.059*	*0.007*	**0.164**	*0.058*	*0.041*	*-0.001*	*0.058*	*0.048*	**0.225**	0.095
1	*-0.030*	*-0.027*	0.133	nan	-0.100	nan	*0.052*	nan	*0.028*	nan
2	-0.077	*-0.014*	*-0.127*	nan	*0.020*	*0.013*	**0.161**	*-0.030*	**0.512**	*0.031*
3	*-0.063*	*-0.016*	**0.192**	*-0.037*	*0.001*	*0.015*	*-0.037*	*0.019*	**0.331**	0.158
4	*-0.048*	*0.002*	*0.064*	*0.006*	*-0.021*	*0.025*	*0.019*	**0.090**	0.165	**0.275**
5	-0.084	*-0.059*	*0.059*	*-0.013*	*-0.026*	-0.129	-0.254	*0.056*	**0.322**	0.166
6	*-0.008*	*-0.008*	**0.183**	*0.030*	-0.188	*-0.040*	**0.213**	0.174	**0.418**	0.173
7	*-0.007*	*-0.006*	*-0.007*	*-0.045*	*0.018*	*0.020*	*-0.019*	*0.013*	**0.158**	0.090
8	*0.031*	*0.064*	**0.214**	*0.023*	*0.013*	*0.024*	*-0.007*	*0.062*	*0.029*	**0.089**
9	*-0.029*	*0.000*	**0.142**	*0.017*	*0.037*	*-0.035*	*-0.018*	**0.127**	-0.153	**0.092**
10	*-0.020*	*0.009*	*0.070*	*-0.006*	*0.028*	*0.003*	*0.026*	**0.084**	**0.366**	0.142

Content Preservation

Prompt	Falcon 7b	Falcon 7b Restr.	Falcon 40b	Falcon 40b Restr.	Llama2 7b	Llama2 7b Restr.	Llama2 13b	Llama2 13b Restr.	Llama2 70b	Llama2 70b Restr.
0	*-0.028*	*0.025*	*0.060*	-0.082	*-0.018*	*0.030*	*0.008*	**0.079**	**0.278**	*-0.001*
1	*-0.012*	*0.057*	*-0.058*	*-0.011*	*0.067*	**0.078**	*-0.049*	*-0.017*	*-0.029*	*0.059*
2	**0.075**	*0.041*	**0.179**	*0.028*	*-0.036*	*-0.055*	*0.051*	-0.153	**0.248**	*0.018*
3	*0.050*	*0.035*	*0.094*	nan	**0.111**	*0.008*	**0.148**	*0.009*	**0.331**	*-0.068*
4	-0.101	*-0.019*	0.129	nan	-0.115	*-0.019*	*0.084*	-0.086	**0.191**	*-0.003*
5	*0.002*	*0.021*	**0.159**	*-0.071*	*0.039*	*-0.053*	*0.005*	-0.104	**0.091**	*-0.026*
6	*0.025*	*-0.036*	**0.074**	*0.057*	*-0.111*	*-0.030*	**0.093**	*-0.021*	**0.124**	0.095
7	*0.020*	*0.009*	*0.057*	*0.027*	*-0.001*	*0.024*	*-0.043*	*0.027*	*-0.001*	*0.044*
8	*0.010*	*-0.055*	**0.124**	*0.062*	*0.312*	0.103	*0.046*	**0.099**	0.115	**0.164**
9	*0.013*	*0.010*	**0.106**	0.094	*-0.022*	**0.078**	*0.064*	*0.055*	**0.270**	*0.065*
10	*-0.005*	*-0.065*	0.202	nan	*-0.010*	nan	*-0.015*	nan	**0.289**	*0.068*

Fluency

Prompt	Falcon 7b	Falcon 7b Restr.	Falcon 40b	Falcon 40b Restr.	Llama2 7b	Llama2 7b Restr.	Llama2 13b	Llama2 13b Restr.	Llama2 70b	Llama2 70b Restr.
0	**0.083**	*-0.004*	0.293	**0.385**	**0.313**	0.245	*-0.058*	**0.356**	**0.484**	0.417
1	*-0.040*	*0.027*	**0.266**	*0.054*	**0.117**	*-0.038*	-0.100	*-0.020*	**0.636**	*0.000*
2	*0.035*	*0.050*	**0.309**	0.188	**0.243**	*-0.017*	**0.248**	0.182	**0.557**	0.333
3	*0.009*	*0.051*	0.134	nan	**0.079**	*-0.006*	**0.151**	*0.037*	0.382	nan
4	*-0.023*	*0.030*	0.223	nan	0.201	nan	**0.137**	*-0.021*	0.494	nan
5	*-0.047*	*0.004*	**0.441**	0.133	**0.310**	*0.069*	**0.350**	0.163	**0.609**	0.193
6	*0.045*	*0.057*	**0.153**	*0.063*	*0.032*	*0.000*	*0.000*	*0.015*	*-0.018*	*0.023*
7	*-0.049*	*0.041*	**0.417**	*0.063*	**0.173**	*0.054*	0.098	**0.192**	**0.610**	0.465
8	*0.060*	**0.117**	**0.198**	0.103	*-0.046*	**0.113**	**0.227**	*0.032*	**0.507**	0.275
9	*0.038*	*0.033*	**0.393**	0.215	0.100	**0.103**	0.115	**0.154**	**0.587**	0.283
10	*-0.044*	*-0.001*	**0.098**	*0.060*	**0.193**	*0.055*	**0.127**	*-0.020*	**0.284**	-0.083

Table 5: Shown are the Spearman rank correlations for style transfer accuracy (top), content preservation (middle), and fluency (bottom) between human evaluations and the fine-tuned LLM's evaluations for individual prompts, comparing unrestricted sampling and sampling restricted to decimals. All *italic correlations* have p>0.05.

and Llama2 we deploy the pre-trained models on our own hardware using Alpa (Zheng et al., 2022).

To investigate the impact of different language model sizes, we use the OPT (Zhang et al., 2022) family of models, as well as Falcon (Almazrouei et al., 2023) and Llama2 (Touvron et al., 2023). The OPT models have demonstrated performance comparable to GPT2 (Radford et al., 2019) and

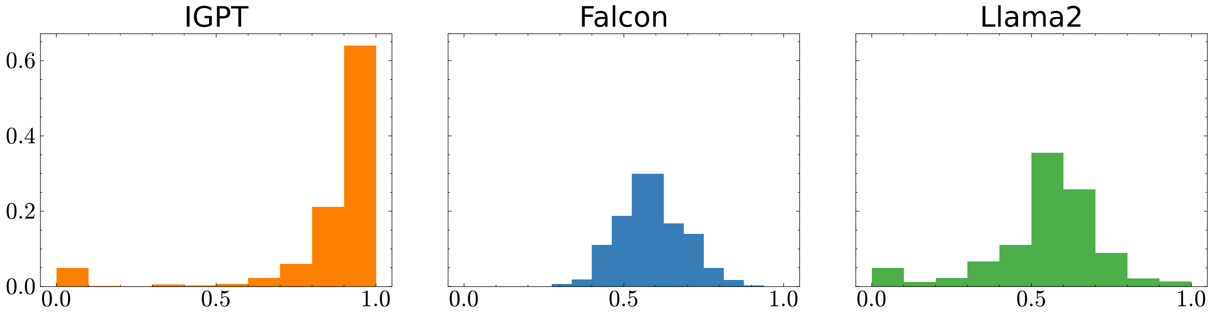


Figure 4: Shown is the distribution for LLM evaluations for InstructGPT (IGPT with 175b parameters), Falcon ("instruct" with 40b parameters), and Llama2 ("chat" with 70b parameters) for the aspect of style transfer accuracy. IGPT returns extreme evaluations, while Falcon and Llama2 are centered around 0.5.

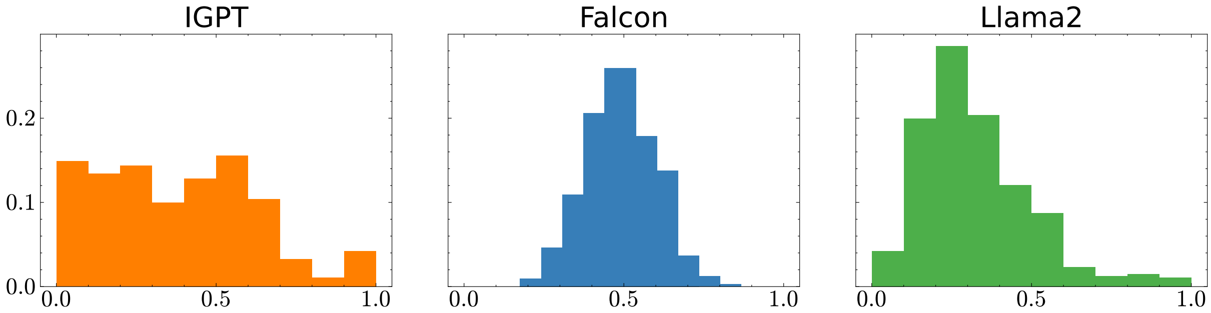


Figure 5: Shown is the distribution for LLM evaluations for InstructGPT (IGPT with 175b parameters), Falcon ("instruct" with 40b parameters), and Llama2 ("chat" with 70b parameters) for the aspect of fluency. IGPT returns almost uniformly distributed evaluations between 0.1 and 1.0. At the same time, Llama2 is centered around 0.5, while Falcon is skewed towards 0.0.

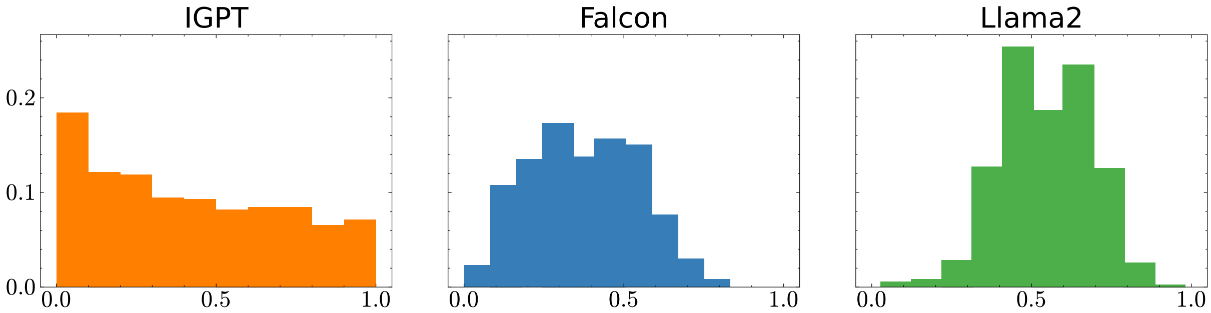


Figure 6: Shown is the Spearman rank correlation of each prompt with human evaluations for InstructGPT (IGPT with 175b parameters), Falcon ("instruct" with 40b parameters), and Llama2 ("chat" with 70b parameters). The horizontal dashed lines indicate the correlation of the prompt ensemble. The ensemble tends to have a higher correlation than individual prompts.

GPT3 (Brown et al., 2020). Most OPT models are freely available for use (except for the largest OPT-175b, which requires a request for access). The OPT models we use have sizes of 125m, 350m, 1.3b, 2.7b, 6.7b, 13b, 30b, 66b, and 175b parameters. We use the 7b and 40b models for Falcon in the "normal" version. For Llama2, we run the models with 7b, 13b, and 70b parameters in the "normal" version.

D. Largest Pre-trained Models

This section presents the results for the largest pre-trained LLMs deployed for TST evaluation.

D.1. Ensembled Prompts

The results are summarized in Table 6 (see also Table 1 for other automated metrics).

Style Transfer Accuracy				
	ARAE	CAAE	DAR	All
OPT-175b	*-0.112*	*-0.052*	*0.007*	*-0.039*
BLO-176b	0.311	*-0.052*	*0.107*	0.118
GPT-175b	0.126	*-0.042*	*0.046*	*0.044*
Fal-7b	*0.049*	*-0.058*	*0.030*	*0.013*
Fal-40b	*0.058*	*0.016*	0.186	0.094
Lla-7b	0.144	*-0.026*	*-0.014*	*0.030*
Lla-13b	0.271	*0.109*	0.215	0.191
Lla-70b	0.350	0.406	0.389	0.393
Content Preservation				
	ARAE	CAAE	DAR	All
OPT-175b	*-0.013*	*-0.085*	*-0.080*	*-0.067*
BLO-176b	*-0.019*	*-0.051*	*-0.050*	*-0.042*
GPT-175b	*0.028*	*0.008*	*0.090*	*0.042*
Fal-7b	*-0.070*	*-0.026*	*-0.083*	*-0.036*
Fal-40b	*0.047*	*0.104*	0.189	0.114
Lla-7b	*-0.009*	*0.000*	*-0.017*	*-0.011*
Lla-13b	*0.002*	*0.027*	*0.055*	*0.051*
Lla-70b	*0.083*	0.464	*0.025*	0.212
Fluency				
	ARAE	CAAE	DAR	All
OPT-175b	*-0.034*	*-0.102*	*0.005*	*-0.058*
BLO-176b	-0.127	*-0.101*	*-0.044*	-0.101
GPT-175b	*0.030*	*0.053*	*-0.015*	*0.030*
Fal-7b	*-0.019*	*0.079*	*-0.034*	*0.016*
Fal-40b	0.218	0.170	0.194	0.200
Lla-7b	*0.107*	0.199	*-0.011*	0.093
Lla-13b	0.242	0.215	0.129	0.205
Lla-70b	0.436	0.540	0.479	0.521

Table 6: Shown are the Spearman rank correlations for style transfer accuracy (top), content preservation (middle), and fluency (bottom) between human evaluations and the mentioned automated metrics, including OPT, BLOOM (BLO), GPT3 (GPT), Falcon (Fal), and Llama2 (Lla). All *italic correlations* have p>0.05.

D.2. LLM Limitations and Failure Modes

D.2.1. Parsable Answers

Table 7 summarizes our findings in terms of parsable answers of the pre-trained LLMs. Most answers are parsable for OPT and BLOOM, where 85.8-93.2% of the answers of BLOOM and about 80.8-93.6% OPT's answers are parsable. GPT3 has considerably higher rates of parsable answers. However, Llama2 with 70b parameters exhibits the highest number of answers that can be parsed despite having fewer parameters. In general, Falcon and Llama models have a greater proportion of parsable answers compared to OPT and BLOOM. Specifically, these models yield at least 90% parsable answers across all three aspects of evaluation.

We can observe that adding the phrase "Result =" as a suffix of the prefix prompt increases the number of parsable answers. Overall, we can see that evaluating text style transfer accuracy is more reliable for OPT, BLOOM, and Llama2 than content preservation and fluency. GPT3 and Falcon have the worst parsing rates for content preservation.

As shown, the largest investigated language models return a numerical score at least 80% of the time. To see the effect of model size, we investigate all available pre-trained model sizes of OPT in Appendix E, demonstrating the increased reliability of bigger models in parsable answers.

	STA	CP	F
OPT	93.6%	88.3%	80.8%
BLOOM	93.2%	89.9%	85.8%
GPT3	96.3%	94.2%	96.8%
Falcon-7b	95.3%	91.2%	97.3%
Falcon-40b	95.9%	90.0%	96.5%
Llama2-7b	96.8%	94.0%	94.6%
Llama2-13b	96.9%	93.4%	95.8%
Llama2-70b	98.1%	95.5%	98.1%

Table 7: Shown is the proportion of answers for the three largest pre-trained LLM evaluation models OPT, BLOOM, GPT3, Falcon, and Llama2 where the answer is parsable to return a score for the aspects of style transfer accuracy (STA), content preservation (CP), and fluency (F).

D.2.2. In-Range Scores

Table 8 summarizes the results of our study on pre-trained LLMs. Llama2-70b has the highest number of in-range scores, closely followed by Falcon-40b, BLOOM, and GPT3. OPT has the least in-range scores. Generally, larger models tend to have more in-range scores. We investigate the impact of LLM size on the number of in-range scores for all available pre-trained OPT models in Appendix A, demonstrating lower in-range scores for smaller models.

E. Smaller Pre-trained Models

This section presents the results for smaller LLM evaluations with different OPT sizes.

E.1. Correlations with Human Evaluations

As highlighted in Section 5.1, ensembling enhances the robustness of our LLM evaluation. Therefore, we exclusively report ensembled correlations in this section. We summarize our findings regarding the correlations of smaller LLM evaluations with human

	STA	CP	F
OPT	94.8%	94.1%	93.4%
BLOOM	98.8%	98.7%	97.1%
GPT3	95.7%	97.4%	97.0%
Falcon-7b	88.8%	92.1%	99.0%
Falcon-40b	99.0%	96.4%	98.9%
Llama2-7b	96.2%	96.6%	98.4%
Llama2-13b	98.8%	98.8%	99.2%
Llama2-70b	99.1%	98.5%	99.2%

Table 8: Shown is the proportion of answers for the largest pre-trained LLM evaluation models OPT, BLOOM, GPT3, Falcon, and Llama2 where the parsed score is within the given range in the prompt for the aspects of style transfer accuracy (STA), content preservation (CP), and fluency (F).

evaluations in Table 9 for style transfer accuracy at the top, content preservation in the middle, and fluency at the bottom.

As mentioned earlier in Section 5, even the largest OPT model with 175 billion parameters exhibits correlations close to zero or slightly negative, with p-values > 0.05 indicating non-reportable correlations for all three evaluation aspects and all investigated TST models (including the combination of their outputs). These results also extend to the smaller LLMs: we also observe correlations close to zero or slightly negative, with p-values > 0.05 for most of the reported correlations on all three evaluation aspects across all investigated TST models.

E.2. Parsable Answers

We summarize our findings for parsable answers of smaller LLMs in Figure 7. Overall, we observe a clear trend: the larger the language model, the more parsable the answers.

For evaluating style transfer accuracy, the fraction of parsable answers increases from approximately 0.75 for the smallest 125m model to 0.95 for the 2.7b model and remains at that level. Our analysis shows that bigger models are not necessarily more reliable.

The evaluation of content preservation exhibits a similar trend to style transfer accuracy, with the fraction of parsable answers increasing from 0.8 for the smallest 125m model to around 0.88 for the 2.7b model and remaining stable at that level.

For fluency, the fraction of parsable answers is highest across almost all model sizes (except for the 1.3b model). The fraction of parsable answers starts at around 0.65 for the smallest 125m model and increases to approximately 0.75 for the largest models, although the trend is not consistent for the 30b to 175b models.

Style Transfer Accuracy				
	ARAE	CAAE	DAR	All
OPT125m	*-0.060*	*-0.049*	*0.100*	*0.004*
OPT350m	*0.076*	*0.121*	*0.008*	0.074
OPT1.3b	*-0.009*	*0.085*	*0.047*	*0.064*
OPT2.7b	*-0.038*	*0.061*	*-0.031*	*0.000*
OPT6.7b	*0.015*	*0.039*	*0.008*	*0.035*
OPT13b	*0.018*	*0.011*	*-0.097*	*-0.017*
OPT30b	*0.094*	*-0.093*	*-0.079*	*-0.029*
OPT66b	*0.051*	*-0.060*	*-0.037*	*-0.016*
OPT175b	*-0.112*	*-0.052*	*0.007*	*-0.039*
Content Preservation				
	ARAE	CAAE	DAR	All
OPT125m	*-0.076*	*-0.091*	*-0.028*	*-0.056*
OPT350m	*-0.092*	*-0.040*	*-0.036*	*-0.047*
OPT1.3b	*0.052*	*-0.010*	*-0.014*	*0.006*
OPT2.7b	*0.043*	*-0.002*	*-0.006*	*0.006*
OPT6.7b	*-0.076*	*-0.077*	-0.129	-0.091
OPT13b	*-0.012*	*-0.005*	*0.019*	*0.002*
OPT30b	*-0.049*	*0.021*	*0.060*	*-0.022*
OPT66b	*0.000*	-0.162	*0.107*	*-0.026*
OPT175b	*-0.013*	*-0.085*	*-0.080*	*-0.067*
Fluency				
	ARAE	CAAE	DAR	All
OPT125m	*-0.020*	*-0.029*	*-0.079*	*0.008*
OPT350m	*0.064*	*0.068*	*0.023*	*0.051*
OPT1.3b	*-0.006*	*-0.054*	-0.133	-0.078
OPT2.7b	*0.029*	*0.031*	*-0.018*	*0.019*
OPT6.7b	*-0.034*	*-0.077*	*-0.039*	*-0.069*
OPT13b	*-0.067*	*-0.109*	*-0.055*	*-0.084*
OPT30b	*-0.111*	-0.165	*-0.053*	-0.141
OPT66b	0.176	*0.061*	*0.114*	0.112
OPT175b	*-0.034*	*-0.102*	*0.005*	*-0.058*

Table 9: Shown are the Spearman rank correlations for style transfer accuracy (top), content preservation (middle), and fluency (bottom) between human evaluations and the LLM evaluations with different model sizes of OPT. All *italic correlations* have p>0.05.

E.3. In-Range Scores

We summarize our findings for in-range scores of smaller LLMs in Figure 8. We observe a similar trend for in-range scores as for parsable answers: smaller LLMs are less reliable and return less in-range scores than larger LLMs. However, the trend exhibits more oscillation compared to parsable answers.

In evaluating style transfer accuracy, the fraction of in-range scores inreases from approximately 0.8 for the smallest 125m model to 0.95 for the largest 175b model. Intermediate model sizes exhibit oscillation around this upward trend.

Content preservation evaluation follows a similar trend as style transfer accuracy, with the fraction of

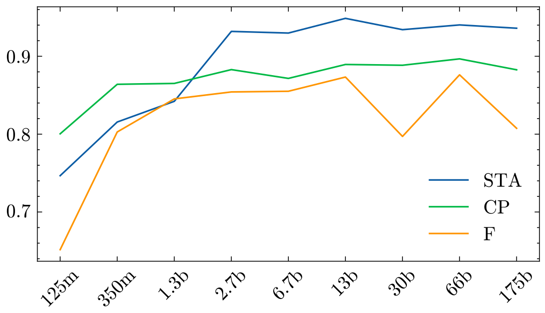


Figure 7: Shown is the fraction of parsable answers for different model sizes of OPT. There is a clear trend that larger models return more parsable answers for all three evaluation aspects of style transfer accuracy (STA), content preservation (CP), and fluency (F).

in-range scores increasing from 0.8 for the smallest 125m model to around 0.95 for the largest 175b model. The oscillation is more pronounced compared to style transfer accuracy.

Fluency evaluation shows the most in-range scores starting from the 350m model. The trend remains consistent with the other evaluation aspects: larger models are more reliable than smaller models. The oscillation is smaller, except for the 30b model.

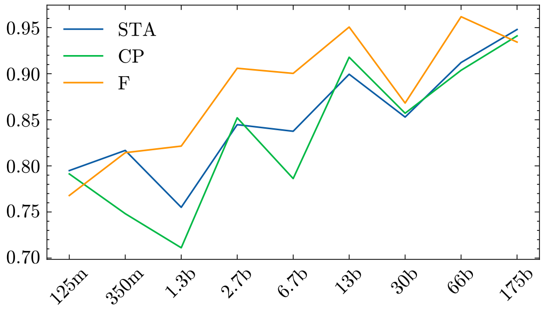


Figure 8: Shown is the fraction of answers where the parsed score is within the given range for different model sizes of OPT. There is a clear trend that larger models return more in-range scores for all three evaluation aspects of style transfer accuracy (STA), content preservation (CP), and fluency (F).

F. Prompts

We present all deployed prompts for style transfer accuracy in Table 10, for content preservation in Table 11, and for fluency in Table 12.

Index	Prompt
0	Here is sentence S1: **{input}** and sentence S2: **{transferred}**. How different is sentence S2 compared to S1 on a scale from 1 (identical styles) to 5 (completely different styles)? Result =
1	Here is sentence S1: **{input}** and sentence S2: **{transferred}**. How different is sentence S2 compared to S1 on a continuous scale from 0 (identical styles) to 1 (completely different styles)? Result =
2	Please evaluate the style transfer intensity between sentence A **{input}** and sentence B **{transferred}** on a scale of 1 to 5, where 1 represents an identical style and 5 represents a completely different style.
3	How different is sentence S1 = **{input}** compared to S2 = **{transferred}** on a scale from 1 (identical styles) to 5 (completely different styles)? Result =
4	How different is the sentence S1 = **{input}** compared to S2 = **{transferred}** for style [positivity] on a scale from 1 (identical styles) to 5 (completely different styles)? Result =
5	The sentence S2 = **{transferred}** is a style transfer of sentence S1 = **{input}**, on a scale from 1 (identical styles) to 5 (completely different styles) evaluate the style transfer intensity between S1 and S2? Result =
6	Here is sentence S1: **{input}**, sentence S2: **{transferred}** and style S3 [sentiment]. How different are S1 and S2 for S3 style on a scale from 1 (identical styles) to 5 (completely different styles)? Result =
7	Here is sentence S1: **{input}**, sentence S2: **{transferred}** and style S3 [sentiment]. How different are S1 and S2 for S3 style on a discrete scale from 1 to 5 where [1 = completely identical styles, 2 = identical styles, 3 = not identical nor different styles, 4 = different styles, 5 = completely different styles]? Result =
8	Here is sentence S1: **{input}** and sentence S2: **{transferred}**. How different is sentence S2 compared to S1 on a discrete scale from 1 to 5 where [1 = completely identical styles, 2 = identical styles, 3 = not identical nor different styles, 4 = different styles, 5 = completely different styles]? Result =
9	Here is sentence S1: **{input}** and sentence S2: **{transferred}**. How different is sentence S2 compared to S1 on a continuous scale from 1 (completely identical styles) to 5 (completely different styles)? Result =
10	How different is the style of sentence S1 = **{input}** compared to S2 = **{transferred}** on a scale from 1 (identical styles) to 5 (completely different styles)? Result =

Table 10: Shown are the prompts to measure style transfer accuracy.

Index	Prompt
0	Here is sentence S1: **{input}** and sentence S2: **{transferred}**. The sentences S1 and S2 have the opposite sentiment but how much does the content change on a scale from 1 (completely different content) to 5 (identical content) on a continuous scale? Result =
1	Here is sentence S1: **{input}** and sentence S2: **{transferred}**. The sentences S1 and S2 have the opposite sentiment but has the content changed on a scale from 1 (completely changed) to 5 (not changed)? Result =
2	Here is sentence S1: **{input}** and sentence S2: **{transferred}**. How different is the topic of sentence S2 compared to S1 on a continuous scale from 1 (completely different topic) to 5 (identical topic)? Result =
3	Please rate the content preservation between the following two sentences on a scale from 1 to 5, ignoring any differences in style or formatting: Sentence 1: **{input}** Sentence 2: **{transferred}** To determine the content preservation between these two sentences, consider only the information conveyed by the sentences and ignore any differences in style or formatting. Based on your evaluation, please provide a rating on a scale from 1 to 5, with 1 being very low content preservation and 5 being very high content preservation.
4	Please evaluate the content preservation between sentence A **{input}** and sentence B **{transferred}** on a scale of 1 to 5, where 1 represents identical content and 5 represents completely different content.
5	How much is the content of sentence S2 **{input}** changed from S1 **{transferred}** on a scale from 1 (completely different content) to 5 (identical content)? Result =
6	How much is the content of sentence S2 **{input}** changed from S1 **{transferred}** neglecting the change in style on a scale from 1 (completely different content) to 5 (identical content)? Result =
7	Here is sentence S1: **{input}** and sentence S2: **{transferred}**. The sentences S1 and S2 have the opposite [sentiment], but has the content changed on a scale from 1 to 5 where 5 (Equivalent), 4 (Roughly equivalent), 3 (Not equivalent but share some details), 2 (Not equivalent but on the same topic) and 1 (Completely dissimilar)? Result =
8	Here is the input sentence S1: **{input}** and transferred sentence S2: **{transferred}**. How much does S2 preserve the content of S2 on a scale from 1 to 3 where 3 (The transferred sentence has the same meaning as the input sentence), 2 (The transferred sentence preserves part of the content meaning of the input sentence), and 1 (The transferred sentence and the input sentence are irrelevant to the content)? Result =
9	Here is the input sentence S1: **{input}** and transferred sentence S2: **{transferred}**. How much does S2 preserve the content of S2 on a scale from 1 (completely changed) to 5 (not changed)? Result =
10	Here is S1: **{input}** and sentence S2: **{transferred}**. How much does S2 preserve the content of S2 on a continuous scale from 0 (completely different topic) to 1 (identical topic)? Result =

Table 11: Shown are the prompts to measure content preservation.

Index	Prompt
0	Here is sentence S1: **{input}** How fluent is the sentence S1 on a continuous scale from 1 (lowest fluency) to 5 (highest fluency)? Result =
1	Please evaluate the fluency, grammar, and naturalness of the following sentence **{input}** on a scale of 1 to 5, where 1 represents poor fluency, grammar, and naturalness and 5 represents excellent fluency, grammar, and naturalness.
2	How fluent is the sentence **{input}** on a continuous scale from 1 (lowest fluency) to 5 (highest fluency)? Result =
3	How fluent is the sentence **{input}** on a continuous scale from 0 (lowest fluency) to 1 (highest fluency)? Result =
4	How fluent is the sentence **{input}** on a scale from 0 to 1 where 0 equals (lowest fluency) and 1 equals (highest fluency)? Result =
5	How fluent is the sentence **{input}** on a scale from 1 to 5 where 5 (perfect), 4 (comprehensible), 3 (somewhat comprehensible), 2 (incomprehensible), and 1 (incomplete or just a fragment)? Result =
6	How are the grammars of the sentence **{input}** on a scale from 1 to 5 where 1 (lowest grammar preserving) and 5 (highest grammar preserving)? Result =
7	On a scale from 1 to 5, rate the fluency and naturalness of sentence S1 **{input}** where 1 (lowest rate) and 5 (highest rate)? Result =
8	On a scale from 1 to 5, how coherent is **{input}** where 1 (lowest coherent) and 5 (highest coherent)? Result =
9	How natural is this sentence S1 **{input}** on a scale from 1 to 5 where 1 (lowest coherent) and 5 (highest coherent)? Result =
10	S1 = **{input}** Rate the fluency of S1 on a scale from 1 (lowest fluency) to 5 (highest fluency).

Table 12: Shown are the prompts to measure fluency.